Available online at www.sciencedirect.com

SCIENCE DIRECT®

Acta Psychologica 117 (2004) 251–274

acta psychologica

www.elsevier.com/locate/actpsy

Actor–Observer differences in realism in confidence and frequency judgments

Carl Martin Allwood *, Marcus Johansson

Department of Psychology, Lund University, Box 213, SE-221 00 Lund, Sweden

Received 1 December 2003; received in revised form 3 June 2004; accepted 4 June 2004
Available online 9 September 2004

Abstract

Taking a social psychological approach to metacognitive judgments, this study analyzed the difference in realism (validity) in confidence and frequency judgments (i.e., estimates of overall accuracy) between one's own and another person's answers to general knowledge questions. Experiment 1 showed that when judging their own answers, compared with another's answers, the participants exhibited higher overconfidence, better ability to discriminate correct from incorrect answers, lower accuracy, and lower confidence. However, the overconfidence effect could be attributable to the lowest level of confidence. Furthermore, when heeding additional information about another's answers the participants showed higher confidence and better discrimination ability. The overconfidence effect of Experiment 1 was not found in Experiment 2. However, the results of Experiment 2 were consistent with Experiment 1 in terms of discrimination ability, confidence, and accuracy. Finally, in both experiments the participants gave lower frequency judgments of their own overall accuracy compared with their frequency judgments of another person's overall accuracy.

PsycINFO classification: 2340
Keywords: Realism; Confidence judgments; Frequency judgments; Social influence; Metacognition

* Corresponding author.
E-mail address: cma@psychology.lu.se (C.M. Allwood).

0001-6918/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.actpsy.2004.06.006

1. Introduction

A common experience is having some *degree of confidence* in the accuracy of a memory or a knowledge assertion. For instance, a person might be more or less confident that a meeting is scheduled to take place tomorrow rather than today, that the stove was turned off, or that Gabarone is the capital of Botswana.

Research on the validity or *realism* in confidence judgments of the correctness of general knowledge assertions has shown that people often demonstrate *overconfidence*; that is, the level of their confidence judgments tends to exceed the level of accuracy (for a review, see McClelland & Bolger, 1994). Explanations of the overconfidence effect include cognitive processing biases (e.g., Griffin & Tversky, 1992; Koriat, Lichtenstein, & Fischhoff, 1980), and methodological and statistical factors (e.g., Erev, Wallsten, & Budescu, 1994; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994).

In contrast to confidence judgments, *frequency judgments* (i.e., estimates of overall accuracy) often show lower levels and better realism or even underestimation (see, e.g., Allwood & Granhag, 1996a; Gigerenzer et al., 1991; Granhag, Strömwall, & Allwood, 2000; Sniezek & Buckley, 1991; Treadwell & Nelson, 1996). Based on this so-called *confidence–frequency effect* (Gigerenzer et al., 1991), it has often been concluded that these two metacognitive judgments differ in terms of the content of their underlying processes. For example, one proposal is that confidence judgments involve information about item content, whereas frequency judgments involve, for example, reflections about oneself, including conceptions of the connection between one's expertise and the demands of the task (Sniezek & Buckley, 1991).

To date, the influence that different social circumstances of people's everyday life settings might impose on the degree of realism has only received scant attention. For instance, people's metacognitive judgments, such as confidence and frequency judgments, not only concern the accuracy of their *own*, but also of *other persons*' memory and knowledge assertions (Allwood & Granhag, 1999; Jost, Kruglanski, & Nelson, 1998). An intriguing question in this context is whether the realism in confidence judgments and frequency judgments of one's own and another person's assertions differ. The general aim of the present study is to improve the understanding of this issue. Below, participants judging their own assertions are called Actors and participants judging another person's assertions are called Observers.

Only a few studies have addressed the difference between Actors' and Observers' realism in confidence judgments (Allwood, 1994; Harvey, Koehler, & Ayton, 1997; Koehler, 1994; Koehler & Harvey, 1997). These studies involve task and confidence-scale asymmetries between Actors and Observers (further outlined below), which are relevant for the interpretation of the outcome of the Actor–Observer difference. One aim of the present study was to analyze the effect of these factors.

Furthermore, we asked whether, and if so how, additional information about another's assertions might influence the Observer's realism. The reason for adding information may be to influence another person's confidence in the assertions. For example, when asserting that "Gabarone is the capital of Botswana" the person might add, "I remember having been told that this is the case" or "I am 90% certain''. The first kind of additional information in the example is henceforth called an *argument*, whereas the latter is referred to as additional information in terms of a *confidence judgment*. The expected effect of these two types of additional information is discussed below.

Finally, we also investigated the Actor–Observer difference in *frequency judgments*. In view of the discrepancy between confidence and frequency judgments (described above) it is of interest to investigate whether the Actor–Observer difference in the realism in frequency judgments differs from that of confidence judgments. As far as we know, only Johansson and Allwood (2004) have investigated Actor–Observer differences in the realism in frequency judgments.

1.1. Actor–Observer differences in the realism of confidence judgments

Why might an Actor–Observer difference in the realism in confidence judgments be expected? To begin with, in addition to their own knowledge, Observers also have some information about the Actor's knowledge. Given that the Observer perceives that the Actor's knowledge differs at least somewhat from his/her own, the Observer, on average, has access to more knowledge than the Actor. In this context it is relevant to consider how the Observer is likely to view the other's (Actor's) knowledge. Here, Nikerson's (1999) general model of how people form conceptions of other persons' knowledge is relevant. Nickerson suggested that people form an impression of what other persons know by using their conception of *their own* knowledge as a starting point, and then adjusting this conception by taking into consideration how the other individual's group affiliations and specific person appear to differ from themselves. He also reviewed research showing that people are likely to believe that other persons share more knowledge with themselves than they in fact do.

The theoretical account offered by Kruger (1999) on the optimistic and pessimistic bias effects is also relevant. This account has some similarities to Nickerson's model but is more specific in that it also takes into account the influence of how the task is perceived. The *optimistic bias* is the effect of people rating themselves as better than the average person (e.g., Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995; Kruger & Dunning, 1999; Svenson, 1981), and vice versa for the *pessimistic bias* (e.g., Klar, Medding, & Sarel, 1996; Kruger, 1999; Van Yperen, 1992). Kruger (1999) suggested that the mechanism behind both these biases can be construed of in terms of ''egocentrism''. For instance, when assessing one's own social ability, an optimistic bias would result when one fails to consider that most people have no trouble getting along with others. Likewise, for the pessimistic bias effect, which is more associated with tasks being perceived as difficult, the focus on one's own lack of ability is insufficiently weighed against others also finding the task difficult.

In line with the idea of Kruger (1999) and Nikerson (1999) that people put most emphasis on themselves relative to others, we assume that people's confidence (as well as frequency) judgments of their own and other's knowledge assertions are influenced by the judges' own knowledge. However, since the knowledge questions used in the present study are likely to be perceived as fairly difficult, we assume that the Observers will show a pessimistic bias and, on average, consider the Actor's knowledge as better than their own. This assumption is supported by Johansson and Allwood (2004), who used a between-subjects design and reported that, on average, across a fairly large number of general knowledge areas, knowledge ratings of one's own knowledge were significantly lower than the ratings of another pair member's knowledge.

Leippe's (1980) notion that some factors may more clearly influence the judge's accuracy level, whereas others may influence the judge's confidence level, is of use for further understanding Actor–Observer differences. In general, higher *accuracy* can be expected if the Observer is allowed to adjust his or her own answer after having considered the Actor's answer. One reason for this is that the Observer, when lacking knowledge, may simply choose to agree with the Actor's answer, which may be based on better knowledge.

In addition to increased *confidence* in more accurate performance, other factors may also increase the Observers' confidence. For example, having information about another person's knowledge in terms of his/her answer could be conceptualized as having access to more, or richer, knowledge compared with only having access to one's own knowledge. This might be a reason to feel more confident in a situation where another's assertion is to be confidence judged. This version of the so-called representational richness hypothesis (see Gill, Swann, & Silvera, 1998) is discussed in more detail below. However, as implied above, the perceived information value of another person's knowledge is likely to depend on how its level (or quality) is rated compared with the quality of the person's own knowledge.

In addition, in a situation where another person's assertion is to be confidence judged, the fact that the Observer is likely to pay more attention to the Actor's answer than to other possible answers might, in line with the Support theory (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994), increase the Observer's confidence in the Actor's answer even more. In short, Support theory assumes that "... likelihood judgment reflects an assessment of the balance of evidence favoring the focal hypothesis [i.e., the selected or provided answer] rather than the alternative hypothesis [i.e., any other possible answer]" (Brenner, Koehler, & Rottenstreich, 2002, p. 490).

1.2. Previous research on Actor–Observer differences in realism in confidence

As will be apparent in the following review, the results of the previous studies (Allwood, 1994; Harvey et al., 1997; Koehler, 1994; Koehler & Harvey, 1997) on the Actor–Observer difference are mixed, and suggest that several variables, including materials, tasks, and confidence scales, might influence this difference.

Allwood (1994) used two-alternative forced choice (2AFC) general knowledge questions (GKQs) and a within-subject design. When performing as Actors, the participants answered the GKQs and gave their confidence judgments on a half-range scale (50–100%). When performing as Observers, a full-range scale (0–100%) was used. As Observers, the participants were instructed that confidence judgments lower than 50% meant that they favored the other answer alternative (i.e., the one not selected by the Actor). Allwood attempted to control for the otherwise potentially confounding factor of the "first-Actor-and-then-Observer" order by letting each pair member answer and confidence judge one set of GKQs that was the same (Old) and one set that was different (New) between the two pair members. The results showed that, as Observers, compared with as Actors, the participants showed higher accuracy, higher confidence, larger absolute deviation between confidence and accuracy (i.e., *calibration*), and higher overconfidence. The results also showed a trend towards better discrimination (*resolution*) for the Observers. An interpretation difficulty of these results is that the half-range scale has been found associated with less overconfidence than the full-range scale (Juslin, Wennerholm, & Olsson, 1999; Ronis & Yates, 1987; Weber & Brewer, 2003).

Harvey et al. (1997), as well as Koehler (1994), and Koehler and Harvey (1997) used the full-range scale for both Actors and Observers. Another similarity between these three studies is that they used a between-subjects design. Harvey et al. (1997, see also Koehler & Harvey, 1997, Exp. 3) used a clinical decision-making task for which participant-psychiatrists (Actors) made treatment recommendations. Both participant-nurses (i.e., Observers) and Actors confidence judged the Actors' recommendations. In contrast to Allwood (1994), Harvey et al. (1997) found that the Actors were more overconfident, and more poorly calibrated than the Observers, both when the Observers did not give any treatment recommendations and when the Observers' recommendations were followed by feedback with respect to their effectiveness. The experiments showed no Actor–Observer difference in resolution.

In Koehler (1994), the participants generating predictions (Exp. 1), social inferences (Exp. 3 and 5–6), and answers to GKQs (Exp. 4), gave lower confidence judgments than the participants who evaluated another participant's proposals. Exp. 4, which was the only experiment using calibration measures, also showed that the generate condition resulted in better calibration and better resolution than the evaluate condition. In Exp. 2 (see also Exp. 3), confidence was higher in a condition in which the participants confidence judged their own choice of answers, compared with when the answers selected by another participant were confidence judged. These last results, hence, pertain to a situation in which a closed set of answer alternatives was used. In addition, the insertion of a delay, encompassing a distractor task, between the generation task and the confidence judgment task resulted in essentially the same level of confidence as for the evaluation condition (Exp. 5).

Koehler and Harvey (1997) let their participants identify the boundary shapes of countries (Exp. 1), and provide quantitative answers to historical events (Exp. 2). Koehler and Harvey reported similar results to those of Allwood (1994) and Koehler (1994, Exp. 4). The only difference was that in Exp. 2 of Koehler and Harvey (1997) no difference in mean confidence resulted between Actors and Observers.

The different studies reviewed above show various differences in design which affect the interpretation of their results. For example, as already noted, a problem in Allwood (1994) was that the participants used a half-range confidence scale when performing as Actors, while as Observers they used a full-range confidence scale. In Harvey et al. (1997), Koehler (1994) and Koehler and Harvey (1997), both Actors and Observers used the full-range scale. A further difference between the studies is that the accuracy level was allowed to differ between Actors and Observers in

Allwood (1994), whereas in the analyses of the other three studies the Observers' accuracy was "determined" by the Actors' responses. A problem when the Observers' accuracy level is held constant in this way is that the lowest end-point of the full-range scale (0%, "Absolutely certain that the answer is incorrect") gives contradictory information with respect to the Observers' accuracy, and thus also makes the Observers' confidence–accuracy relationship difficult to interpret.

Another difference between the studies is that while Allwood (1994) used 2AFC GKQs for both Actors and Observers, most of the experiments of the other studies let the Actors generate the answers that they and the Observers confidence judged. In other words, in Allwood (1994) a "recognition" task was used both for Actors and Observers. In the other studies the Actors carried out a "recall" task and the Observers a recognition task. The design of the task used by Harvey et al. (1997), Koehler (1994), and Koehler and Harvey (1997) is a reasonable instance of a situation where Observers evaluate an Actor's assertions, and is useful in order to examine whether considerations of other response alternatives reduce the level of confidence. However, it is still of interest to analyze to what extent the difference in tasks is the cause of the reported differences between Actors and Observers. It is of relevance that stronger confidence–accuracy relations have been found for recall tasks than for recognition tasks (e.g., Robinson & Johnson, 1996; Robinson, Johnson, & Robertson, 2000). In addition, recall and recognition tasks "... differ in terms of retrieval cues and the type of responses they require" (Yonelinas, 2002, p. 450). Consequently, it seems reasonable to suppose that the confidence judgment processes between the two tasks might behave differently.

Given these task and confidence-scale asymmetries between Actors and Observers it is of interest to examine whether the above Actor–Observer differences for GKQs remain in a situation where both the task and the confidence scales are the same for Actors and Observers. For this reason, the present study used 2AFC GKQs and a half-range confidence scale for both Actors and Observers.

Based on the reasoning in the previous section, and because the previous studies using GKQs showed similar results in spite of their methodological differences, we predicted that the participants would show both higher accuracy and confidence when performing as Observers than as Actors. Furthermore, we expected that when the participants performed as Observers they would show worse calibration and higher overconfidence than as Actors (Allwood, 1994; Koehler, 1994, Exp. 4; Koehler & Harvey, 1997, Exp. 1 & 2). We did not pose any hypothesis for resolution due to the inconsistent results of the previous research reviewed above.

1.3. Observers' heeding additional information about Actors' choice

In order to examine how arguments and confidence judgments, as two types of *additional information* about the Actors' answers, influenced Observers' realism we let the participants, when performing as Observers, heed the Actors' arguments for, and/or stated confidence in, some of the Actors' provided answers.

On a general level, similar to the effect of information in terms of the answer alternative selected by the Actor, these two types of additional information might increase the Observer's confidence due to an increase in the amount of information (more information leading to higher confidence). In addition, the content of the provided information may be decisive.

The expectation of the effect of the *amount* of information is inspired by Gill et al.'s (1998) representational richness hypothesis, which states that the richness of people's representation (in terms of amounts of information) of others will contribute to an increase in their confidence. Our version of Gill et al.'s hypothesis states that having more information about the Actor's answers will result in higher confidence, compared with having less information. This hypothesis is also compatible with Support theory (e.g., Tversky & Koehler, 1994) in that more information about the Actor's answer will increase its salience.

Argument information. Koriat et al. (1980) assumed that people are biased in favoring positive rather than negative evidence and found that generating arguments *against* but not *for* the chosen answers to GKQs resulted in somewhat improved realism. However, this result has not been possible to replicate in later studies (Allwood & Granhag, 1996b; Fischhoff & MacGregor, 1982). Furthermore, Allwood and Granhag (1996b) found that arguments provided by the researchers *against* the participants' chosen answers did not improve the realism, although independent judges rated these arguments to be stronger than those generated by the participants.

However, in line with Support theory (e.g., Tversky & Koehler, 1994), a possibility is that the Actor's arguments *for* the chosen answers would increase the Observer's confidence by means of further enhancing the salience of the Actor's chosen answer, relative to the other answer alternative.

Confidence information. While arguments provide information about *why* the answer was chosen, confidence judgments supply information about *the extent* to which one might deem the answer valid. Given this, one possibility is that the Observer might adjust his or her confidence in the direction of the Actor's confidence judgment. However, given that Actors and Observers from the beginning do not, on average, differ in the extremity of their confidence, the net effect of such an averaging process can be expected to be negligible. For this reason we predicted that providing information about the Actor's confidence would not (apart from what has been said above) have any effect on the Observer's confidence.

1.4. Actor–Observer differences in realism in frequency judgments

With respect to the participants' frequency judgments of their own and another person's total number of correctly answered questions, we expected that both these frequency judgments would show the so-called confidence–frequency discrepancy mentioned above. In addition, in line with previous results from research by Johansson and Allwood (2004), we expected that the frequency judgments of another person's answers would show a higher level than the corresponding judgment of one's own answers. Using a between-subjects design and GKQs, Johansson and Allwood (2004, Exp. 1) found that the Actors gave lower frequency judgments than the Observers. This difference is consistent with the pessimistic bias (Klar et al., 1996; Kruger, 1999).

In Exp. 1 of the present study, we further investigated an order effect found in Exp. 2 by Johansson and Allwood (2004). In the latter experiment each participant made individual frequency judgments of both one's own accuracy and another pair member's accuracy. The order of the two judgments was counterbalanced, and between the two judgments there was a time interval filled with a task that involved taking a stand with respect to the correctness of each answer. The results clearly showed that the order of judging one's own accuracy *first*, and *then* the other's accuracy, was associated with lower and more realistic frequency judgments of both one's own and the other's accuracy, and vice versa for the inverse of this order. This order effect can be interpreted as an instance of anchoring or assimilation of a judgment toward a previously considered standard (Tversky & Kahneman, 1974). In Exp. 1 of the present study, in which the two frequency judgments were carried out after one another without any filler task, we expected that no order effect would be found. This expectation is consistent with a proposition by Mussweiler and Neumann (2000) suggesting that if a judgment is attributable to a previously considered standard (or judgment), the standard will be viewed as a potential source of contamination, and consequently contrasted away from.

1.5. Calibration methodology

The present study used *calibration methodology* (see Lichtenstein, Fischhoff, & Phillips, 1982) which provides information about distinct aspects of the relation between confidence and accuracy. More specifically, we used three measures of realism: calibration, over/underconfidence (henceforth, *overconfidence*), and resolution (see Appendix A for equations), derived from the Brier (1950) score formula (Murphy, 1973). In short, *calibration* picks up the goodness of fit between confidence and accuracy in terms of the (squared) difference between the level of the confidence judgments and the accuracy in each of a number of confidence classes (e.g., 50–59% ... 90–99%, and 100%). The only difference between calibration and overconfidence is that the latter provides a directional measure of realism in confidence judgments. Loosely speaking, *resolution* reflects one's ability to differentiate between correct and incorrect answers. It should be noted that for calibration and overconfidence a lower score reflects better realism than a higher score, while a higher resolution score denotes better realism than a lower score.

2. Experiment 1

For Exp. 1, we hypothesized that the participants would show worse calibration, higher overconfidence, and higher accuracy and confidence when performing as Observers than as Actors. Furthermore, for the two types of additional information, we hypothesized that in comparison with the base line, where no additional information was provided, the Actor's arguments but not confidence judgments would increase the Observers' confidence judgments. Consequently, the Observers who were provided with both arguments *and* confidence judgments were predicted to show the same level of confidence as Observers who heeded arguments only. Note that the alternative prediction, building on the representational richness hypothesis, is that the Actor's argument *and* confidence would increase the Observer's confidence more than heeding the Actor's argument *or* confidence only, which in turn would result in higher confidence than heeding none of these. Finally, with respect to Actor–Observer differences in the frequency judgments, we expected that the participants would show lower frequency judgments of their own, as compared with the other's, total number of correct answers.

2.1. Method

2.1.1. Participants

One hundred and six university students (62 women and 44 men, M age = 25, SD = 4, ranging from 19 to 40 years) from Lund University, Sweden, participated in the experiment. Each participant was awarded approximately US$7 for partaking. Two participants (both female) were defined as outliers (criteria: $z > \pm 3.29$, $p < .001$; see Tabachnick & Fidell, 2000, p. 67) and were omitted in the analysis, leaving 104 participants in the study.

2.1.2. Materials

One hundred and twenty 2AFC GKQs were used in the experiment. The GKQs were answered by use of paper and pencil. For each GKQ one of the answer alternatives was correct. The GKQs covered a multitude of subject areas, such as geography, history, literature, and politics.

The GKQs were divided into six equally sized question sets. Next, the question sets were randomized to participants with the following constraints. For each pair of participants (further outlined below), each pair member received four question sets. Between the two members in each pair, two of the sets differed (New), while the other two sets were the same for both members (Old). The question sets were sequenced so that the first and third sets differed between the pair members. This constraint created two New–Old question set combinations (henceforth referred to as the first and second New–Old question set combination, respectively). This New/Old manipulation was used in order to control for the otherwise possible confounding of each participant first performing as an Actor and then as an Observer, that is, controlling for Actors and Observers having encountered the same question once and twice, respectively.

2.1.3. Design

The experiment included three conditions called Argument–Confidence ($n = 36$), Argument ($n = 34$), and Confidence ($n = 34$). Each condition comprised two phases, called Actor and Observer, always administered in this order.

The participants were matched into pairs using the criteria of same gender, similar age and not being acquainted. The intention behind the use of these matching criteria was to restrict possible effects of stereotypes of gender and age categories. The number of male and female pairs was approximately equal between conditions. Each pair was randomly assigned to one of the three conditions.

When partaking as Observers, each participant either heeded the additional information about the Actor's selected answer (that is, the Actor's argument and/or confidence) in the first or in the second New–Old question set combination. The difference between these two response orders constituted the Heed-order factor. That is, the Observers either heeded the Actor's argument and/or confidence for the first question set combination and no such information for the second or vice versa. Each pair in the three conditions was randomly allocated one of these two heed-orders.

2.1.4. Procedure

The participants took part two at a time, separated from each other by a non-transparent screen for the whole experiment and did not communicate.

Actor phase. In all three conditions, all participants first participated as Actors and answered the 80 GKQs. Directly after having answered a question a confidence judgment of the answer was given, using the half-range scale ranging from 50% to 100%. 50% was defined as "Guessing" and 100% as "Absolutely certain" that the chosen answer alternative was correct.

After having answered and confidence judged each question, the Actors in the *Argument–Confidence* condition and the *Argument* condition provided arguments for their own answers to the first or the second New–Old question set combination. The participants were instructed to write down what they regarded as their strongest argument for each of their answers. The written instructions contained some examples of arguments that the participants could use if they could not come up with a stronger argument of their own. For example, a participant might state: "It seems logical", or "It feels right", that the chosen answer alternative is correct. The participants were not permitted to change their answer when they were formulating an argument.

Observer phase. As Observers, the participants made confidence judgments of each of the Actor's 80 answers, using the same scale as before. The Observers' confidence judgments pertained to their own degree of confidence in the correctness of the answer provided by the Actor.

If the Observer thought that the Actor's chosen answer alternative was erroneous, s/he was instructed to select (circle) the other of the two alternatives and confidence judge that alternative, using the 50–100% scale. To make the dissimilar answers on the same questions distinguishable between pair members, one pair member was equipped with a green pencil, and the other pair member with a blue pencil.

For one of the two New–Old question set combinations, the Observers were instructed to heed, for each question (depending on the condition), either the Actor's argument for and/or confidence in his or her answer before giving a confidence judgment. More specifically, in the *Argument–Confidence* condition the Observer heeded both the Actor's argument and confidence. In the *Argument* condition only the Actor's argument was heeded and in the *Confidence* condition only the Actor's confidence judgment was heeded. In the Argument condition the Observers did not have access to the Actors' confidence judgments.

Frequency judgments. Lastly, each participant made a frequency judgment of their own and of the other participant's accuracy by stating *how many GKQs* out of the 80 had been correctly answered. Either the participants frequency judged their own accuracy first and then the other's accuracy, or vice versa. These two orders (Own-Other's and Other's-Own) of the two frequency judgments were randomly counterbalanced and equally distributed across the Argument–Confidence, Argument, and Confidence conditions.

2.2. Results

We first evaluate the effect of first performing as an Actor and then as an Observer. Second, we report on the analysis of the Actor–Observer difference in realism in confidence judgments in the situation where the Observers did not have access to any additional information (No-heed) about the Actors' choice of answers. Third, for the Observer phase, we report on comparisons between the No-heed and Heed items; that is, we evaluate the effect of the different types of information given to the Observers for one of the two New–Old question set combinations. Lastly, we report on the results concerning the Actor–Observer difference in realism in frequency judgments.

Square root transformations on *calibration*, *overconfidence*, and *resolution* were used in the statistical analyses (see Tabachnick & Fidell, 2000) but not in tables and figures or means given in the text. In Table 1 the means (and standard deviations) of these five dependent measures are presented.

2.2.1. Evaluation of the Actor–Observer order

Initially, we evaluated the effect of first performing as an Actor and then as an Observer. Using the Observers' data we compared the GKQs that the participants encountered for the first time as Observers, that is, only once (New), with those GKQs that they also encountered as Actors, that is, questions they encountered for a second time as an Observer (Old). In brief, the results of paired samples *t*-tests showed no significant differences. Hence, as Observers the participants were not more/less realistic, accurate, or confident about Old compared with New GKQs. Below, the New and Old items were collapsed within each New–Old question set combination.

Another possible consequence of an order effect is a "warm-up" effect. In this context, we compared the *accuracy* for the first 50% ($M = .663$) of the GKQs in the Actor phase with the accuracy of the last 50% ($M = .654$). Since the mean accuracy did not increase in the second half of the GKQs, there was thus no evidence of a warm-up effect.

2.2.2. Actor vs. Observer

The calibration graph in Fig. 1 shows the relation between confidence and accuracy, plotted over six confidence classes: 50–59% ... 90–99%, and 100%. The diagonal refers to perfect calibration.

Table 1
Experiment 1: Means and standard deviations (SD) for the realism in Actors' and Observers' confidence judgments for each condition

Measure	Condition								
	Argument–Confidence			Argument			Confidence		
	Actor	Observer		Actor	Observer		Actor	Observer	
		No-heed	Heed		No-heed	Heed		No-heed	Heed
Calibration	.035 (.023)	.036 (.023)	.029 (.017)	.037 (.022)	.040 (.024)	.047 (.029)	.040 (.023)	.038 (.024)	.031 (.017)
Overconfidence	.034 (.085)	.018 (.106)	.031 (.063)	.051 (.071)	.037 (.096)	.060 (.086)	.055 (.088)	.017 (.091)	.019 (.073)
Resolution	.034 (.012)	.029 (.015)	.040 (.021)	.045 (.017)	.042 (.024)	.041 (.019)	.035 (.014)	.033 (.021)	.038 (.021)
Accuracy	.659 (.091)	.705 (.090)	.705 (.089)	.676 (.077)	.717 (.069)	.726 (.081)	.640 (.081)	.689 (.075)	.698 (.098)
Confidence	.693 (.056)	.723 (.057)	.735 (.063)	.727 (.064)	.754 (.076)	.787 (.058)	.695 (.071)	.706 (.080)	.717 (.083)

Note: For the Argument–Confidence condition, $n = 36$, and for the Argument condition and Confidence condition, $n = 34$, respectively. For Observer, the No-heed columns refer to the question set combination for which no additional information was provided and the Heed columns refer to the question set combination for which the other's argument and/or confidence judgment was heeded.

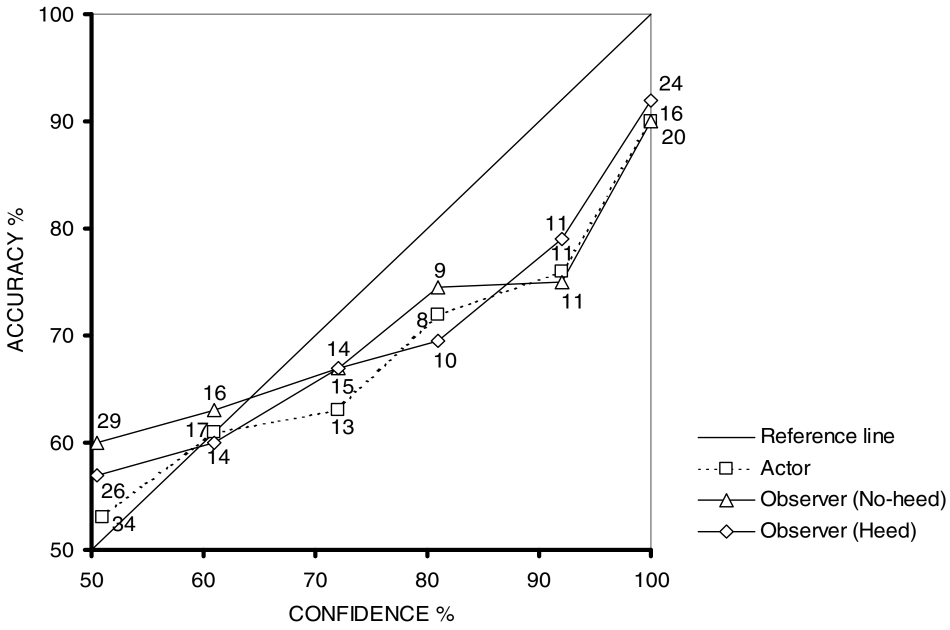

Fig. 1. Experiment 1: Calibration curves for the Actors' confidence judgments and the Observers' confidence judgments for no additional information considered (No-heed) and additional information considered (Heed), respectively, collapsed over the three experimental conditions. The figures in the graph show the percentages of all items for a curve in each confidence class.

As the calibration curves in Fig. 1 show, the difference between the Actor and Observer phases appear largest for the 50–59% confidence class. It is noteworthy that in this confidence class, the participants show fairly good calibration when performing as Actors, while as Observers they show underconfidence, especially when not provided any additional information (No-heed). At the higher confidence levels, the differences between the Actor and Observer curves are less obvious. Fig. 1 also provides the relative frequencies of the confidence judgments at each confidence level. The general pattern shows that the highest and lowest confidence levels were used most frequently, and that the participants used the 50–59% confidence class to a greater extent and the 100% level to a lesser extent when performing as Actors than as Observers.

In order to analyze the Actor–Observer difference in realism in confidence judgments when the participants, as Observers, were not provided with any additional information (No-heed), a paired samples *t*-test was run for each of the five dependent measures (calibration, overconfidence, resolution, accuracy and confidence). The Actor phase consisted of the mean of the two New–Old question set combinations in this phase, while the Observer phase consisted of the New–Old question set combination for which no additional information was heeded. The means and standard deviations are given in Table 1 for both phases (i.e., Actors and Observers) and each condition.

The results showed that when performing as Actors the participants exhibited higher *overconfidence* ($M_{\text{Actor}} = .047$; $M_{\text{Observer}} = .024$), $t(103) = 2.45$, $p = .016$, better *resolution* ($M_{\text{Actor}} = .038$; $M_{\text{Observer}} = .035$), $t(103) = 2.09$, $p = .039$, lower *accuracy* ($M_{\text{Actor}} = .658$; $M_{\text{Observer}} = .704$), $t(103) = -5.05$, $p < .001$, and lower *confidence* ($M_{\text{Actor}} = .705$; $M_{\text{Observer}} = .727$), $t(103) = -5.22$, $p < .001$, than when performing as Observers.

With the intention of evaluating the contribution of the particularly large difference residing at the lowest confidence class (50–59%; see Fig. 1) to the Actor–Observer differences reported above, we excluded (for both phases) the items associated with confidence judgments at that confidence class. When excluding these items the difference in *overconfidence* did not recur ($p > .1$), while the Actor–Observer difference for resolution, accuracy, and confidence remained (all p-values <.05).

2.2.3. Observer phase: no additional information vs. additional information

Next, we carried out a 2 (No-heed/Heed) × 3 (Condition) repeated measures ANOVA on each of the five dependent variables in order to evaluate the effect of additional information on the participants' realism in their confidence judgments when performing as Observers. Note that possible main effects of the Condition factor are of no immediate interest to our hypotheses. The levels of the No-heed/Heed factor refer to the questions for which the participants, as Observers, did not have access to any additional information (No-heed) and the questions for which they had access to the Actors' arguments and/or confidence judgments (Heed). The three levels of the Condition factor were the Argument–Confidence condition, the Argument condition, and the Confidence condition, respectively.

The results showed a significant main effect of the No-heed/Heed factor for *resolution*, $F(1,101) = 4.07$, $p = .046$, and for *confidence*, $F(1,101) = 14.26$, $p < .001$, indicating that better *resolution* ($M_{\text{No-heed}} = .035$; $M_{\text{Heed}} = .040$) and higher *confidence* ($M_{\text{No-heed}} = .727$; $M_{\text{Heed}} = .746$), respectively, were associated with heeding, compared with not heeding, additional information provided by the Actor. Although Table 1 shows that the effect for *resolution* was driven by the Argument–Confidence and the Confidence conditions, the fact that no interaction effect resulted suggests that the *type* of information heeded was inconsequential to the occurrence of improved resolution.

2.2.4. Frequency judgments

In order to evaluate the level of the frequency judgments of one's own and the other's total number of correctly answered questions, and the possible presence of an order effect of first making a frequency judgment of one's own and then the other pair member's accuracy (or vice versa), a 2 (Frequency judgment: Own vs. Other's) × 2 (Order: Own-Other's vs. Other's-Own) ANOVA with repeated measures on the first factor was computed. The means (and standard deviations) for each condition and phase are given in Table 2.

A significant effect of the Frequency judgment factor, $F(1,101) = 7.73$, $p = .006$, showed that the participants' frequency judgments of their *own* accuracy

Table 2
Experiment 1: Means and standard deviations (SD) for frequency judgments and realism in frequency judgments of one's own accuracy and the other's accuracy for two orders of giving these judgments (frequency judgment order condition)

Target	Frequency judgment order condition			
	Own-Other's		Other's-Own	
	Frequency j.	Realism	Frequency j.	Realism
Own accuracy	.578 (.188)	−.086 (.168)	.620 (.188)	−.031 (.164)
Other's accuracy	.624 (.155)	−.041 (.148)	.638 (.176)	−.014 (.183)

Note: $n = 52$ in the Own-Other's condition, and $n = 51$ in the Other's-Own condition.
Frequency j. = frequency judgment/80, and realism = frequency judgment − actual accuracy.

($M = .599$; SD = .188) were significantly lower than their frequency judgments of the *other* pair member's accuracy ($M = .631$; SD = .165). (As indicated by the df, one participant did not carry out both frequency judgments.)

As predicted, the Order factor did not result in a significant effect, neither for the frequency judgments, nor for the corresponding ANOVA that evaluated the order effect for realism in the frequency judgments (frequency judgment minus accuracy); nor did significant Frequency judgment × Order interactions result.

In terms of realism in frequency judgments the participants showed greater *underestimation* in their frequency judgment of their *own* accuracy than in their frequency judgment of the *other's* accuracy, $F(1,101) = 5.92$, $p = .017$. Here, the difference between the frequency judgments of one's own total accuracy and one's actual accuracy was $M = -.059$ (SD = .168) and the corresponding score for the other's accuracy was $M = -.028$ (SD = .166), showing that both judgments were associated with underestimation. In addition, a one-sample *t*-test (test value = 0, i.e., perfect realism in frequency judgment), measuring the realism in the frequency judgments of the participant's *own* accuracy and of the *other's* accuracy respectively, showed that significant *underestimation* was associated with the frequency judgments of one's *own* accuracy, $t(102) = -3.55$, $p < .001$. In contrast, the realism in the frequency judgments of the *other's* accuracy did not differ significantly from zero, $t(103) = -1.70$, $p = .091$.

2.3. Discussion

In sum, the results of Exp. 1 showed that the participants exhibited higher overconfidence, better resolution, lower accuracy, and lower confidence as Actors than as Observers. The overconfidence effect did not remain when the lowest confidence class, 50–59%, was removed. Furthermore, the participants were found to give lower and less realistic frequency judgments of their own accuracy, compared with the other pair member's accuracy. However, no order effect of the frequency judgments resulted.

In Exp. 1 the participants first performed as Actors and then as Observers, and thereby they had had the experience of performing as an Actor when performing as an Observer. In spite of the fact that the two investigated indicators of an order effect (i.e., the presence, or absence, of an increase in accuracy within the Actor phase, and a possible difference in results between Old and New items) were found unreliable, the skeptic might still argue that there is room for interpreting the results of Exp. 1 in terms of a *general* order effect. In order to avoid this possible order effect we used a between-subjects design in Exp. 2.

3. Experiment 2

Exp. 2 examined the general order effect explanation by letting the participants perform either as an Actor *or* as an Observer only. That is, the Actor and Observer roles were between-subjects. The design was similar to Koehler's (1994) and Koehler and Harvey's (1997) study, but differed in that, equivalent to Exp. 1 in the present study, (a) we used 2AFC GKQs for both Actors and Observers, and (b) both Actors and Observers used the half-range confidence scale. Consistent with the results of Exp. 1, we predicted that the Actors would show higher overconfidence, better resolution, lower accuracy, and lower confidence than the Observers. Finally, we expected the frequency judgments of one's own answers to be more pessimistic (i.e., lower) than the frequency judgments of the other's answers.

3.1. Method

3.1.1. Participants

Forty-eight university students (34 women and 14 men, *M* age = 22, SD = 4, ranging from 19 to 42 years) from Lund University, Sweden, participated in the experiment.

3.1.2. Materials and design

Ninety of the 2AFC GKQs used in Exp. 1 were used in Exp. 2. As in Exp. 1, the GKQs were answered, and the answers confidence judged, by means of paper and pencil.

The experiment consisted of two between-subjects conditions, called Actor and Observer. Following the same criteria as in Exp. 1, the participants were matched into pairs. Thus, the main difference, compared with Exp. 1, was that in Exp. 2 each participant took part either as an Actor *or* as an Observer.

3.1.3. Procedure

Actor condition. Here, the participants answered each GKQ by selecting one of the two answer alternatives and directly after having selected (circled) an answer they made a confidence judgment (50–100%) of the selected answer. These confidence judgments were written on a separate response sheet.

Next, each Actor frequency judged his or her own performance by assessing how many of the 90 GKQs they had answered correctly. The frequency judgment was given on a separate sheet.

Finally, the participants in the Actor phase were informed about the aim to investigate the difference between confidence judgments of one's own performance and another's performance. This information was provided for the reason that the Actors were to hand over their answer sheets to the participants who were to perform as Observers. Before the designated Observer arrived, the Actor put the answer sheet, but not the sheets that contained the confidence judgments or the frequency judgment, in an envelope. This was done in order to prevent the Observers to embark on examining the Actors' answers to the GKQs before the experimenter gave further instructions. Before leaving, the Actors handed over the envelope containing the answer sheet to the arriving Observers.

Observer condition. First, each Observer received the envelope with the Actor's answer sheet. Next, each Observer was instructed to confidence judge the Actor's answers, using the 50–100% scale. The Observers were informed that the person who handed over the sheet had answered the GKQs in an attempt to choose the correct answers. Furthermore, equivalent to Exp. 1, the Observers were informed that they could select the other of the two answer alternatives and confidence judge that alternative (also using the 50–100% scale) if they thought that the alternative selected by the other person was erroneous. (The Observers used color pencils that differed from the Actors'.) Lastly, each Observer frequency judged the Actor's actual accuracy.

3.2. Results

Fig. 2 shows the calibration curve for the Actor condition and the Observer condition, respectively. Although no clear overall difference is discernable between the two curves, it is noteworthy that the Actors used the lower confidence classes (50–59%, and 60–69%) more frequently than the Observers, who in turn used the 100% level more often than the Actors.

In Table 3, the means (and standard deviations) are given for each of the five dependent measures (calibration, overconfidence, resolution, accuracy, and confidence), for each condition (Actor and Observer). In order to examine the Actor–Observer differences, a paired samples t-test was run for each measure.

The results showed that neither *calibration* nor *overconfidence* reached significance (p-values > .14). Although only close to significance, the difference for *resolution* was in the predicted direction, $t(23) = 2.02$, $p = .054$. In line with our hypotheses and consistent with Exp. 1, the Actors showed significantly lower *accuracy*, $t(23) = -2.96$, $p = .007$, and lower *confidence*, $t(23) = -2.98$, $p = .007$, than the Observers did.

Also in line with Exp. 1, the frequency judgments (divided by the 90 GKQs) given by the Actors ($M = .484$; SD = .203) were significantly lower than those given by the Observers ($M = .691$; SD = .194), $t(23) = -4.54$, $p < .001$. Furthermore, one-sample t-tests (test value = 0) showed that in terms of realism in frequency judgments (frequency judgment minus actual accuracy) the Actors ($M = -.125$; SD = .192) were associated with significant *underestimation* of their own actual accuracy, $t(23) = -3.21$, $p = .004$. In contrast, the Observers ($M = .081$; SD = .191) were associated with significant *overestimation* of the Actors' actual accuracy, $t(23) = 2.08$, $p = .049$.

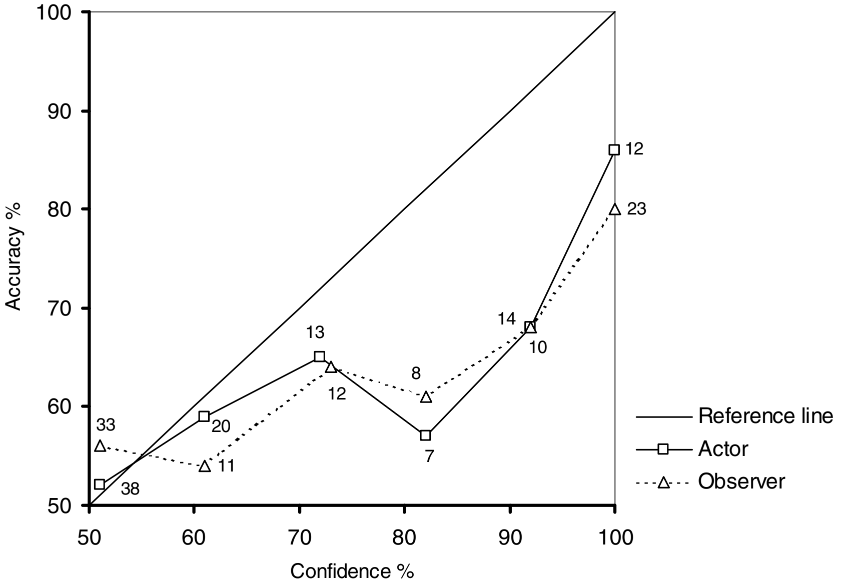

Fig. 2. Experiment 2: Calibration curves for the Actor and Observer conditions. The figures in the graph show the percentages of all items for a curve in each confidence class.

Table 3
Experiment 2: Means and standard deviation (SD) for each of the five dependent measures of Actors and Observers, respectively

Measure	Condition	
	Actor	Observer
Calibration	.029 (.019)	.037 (.020)
Overconfidence	.067 (.073)	.096 (.071)
Resolution	.023 (.010)	.018 (.011)
Accuracy	.611 (.066)	.643 (.069)
Confidence	.677 (.069)	.739 (.072)

3.3. Discussion

In line with our hypotheses, the Actors showed lower accuracy and lower confidence than the Observers. Although resolution did not reach significance, the results were in the predicted direction in that the Actors showed a trend toward better resolution in comparison with the Observers. However, the results showed no significant difference in overconfidence between the Actors and the Observers.

The results of the Actors giving lower frequency judgments than the Observers and the Actors' frequency judgments being associated with underestimation were also in line with our hypotheses. It is noteworthy that the mean value of the Actors' frequency judgments was below the level that would be expected by random performance. Rather than giving realistic frequency judgments the Observers exhibited overestimation in their frequency judgments of the Actors' actual accuracy.

4. General discussion

Taking a social psychological approach to metacognitive judgments, the current study investigated the Actor–Observer difference in the realism in confidence and frequency judgments. For confidence judgments the study examined how different asymmetries in the conditions for Actors and Observers might have influenced the results in previous research. We also investigated the effect of additional information about the Actor's choice of answers on the realism in the Observers' confidence judgments. The additional information was the Actors' arguments for and/or their confidence in their answers. Lastly, the study examined the Actor–Observer difference in realism in the frequency judgments. Below, we summarize and discuss the results pertaining to each of these three issues.

Actor–Observer differences. Our results did not support those reported in previous calibration research on the Actor–Observer difference in realism in confidence judgments (Allwood, 1994; Koehler, 1994, Exp. 4; Koehler & Harvey, 1997, Exp. 1 and 2). Strikingly, in Exp. 1 our results showed that as Observers the participants exhibited *lower* overconfidence than as Actors, whereas the opposite pattern was found in the previous studies. Moreover, when we excluded the 50–59% confidence class in Exp. 1, the difference in overconfidence was eliminated completely. In Exp. 2 we found no significant Actor–Observer difference in realism in confidence judgments. The results showed that the simultaneous increase in accuracy and confidence had the effect to counteract an increase in overconfidence. Taken together, the results of the present study suggest that the task and confidence-scale asymmetries of previous calibration research might have been what produced the Actor–Observer differences in the realism in confidence judgments.

The results of both Exp. 1 and 2 of the present study showed that the Observers exhibited higher accuracy than the Actors. Allwood (1994) reported the same result. A possible explanation of the Observers' higher accuracy pertains to a difference in the response processes when performing as an Actor and as an Observer. When performing as Actors the participants, for each GKQ, first provided an answer and then a confidence judgment of that answer, while as Observers the participants were not required to actively choose an answer alternative to the same extent. For this reason, when performing as Observers, the participants might have freed the processing resources necessary to boost accuracy. In addition, particularly in situations where the Observers were uncertain (i.e., primarily the lowest confidence class), they may have capitalized on the information that the selected answer was the other person's best choice, that is, that s/he had some reason for choosing *A* over *B*, so to speak.

Furthermore, the higher level of confidence in the Observer condition than in the Actor condition is in line with Allwood (1994), Koehler (1994, Exp. 4) and Koehler and Harvey (1997, Exp. 1), and is consistent with the representational richness hypothesis and Support theory (e.g., Tversky & Koehler, 1994).

Observers' heeding of additional information. In Exp. 1 the participants, as Observers, exhibited better resolution and higher confidence when heeding, compared with when not heeding, additional information provided by the Actor. However, we found no interaction effect for resolution or for confidence. This suggests that the type (argument *or* confidence) and, *to some extent*, the amount (argument *and* confidence vs. argument *or* confidence) of additional information heeded were inconsequential. In sum, these results challenge the hypothesis that the type of information about another person's assertions might affect the level of overconfidence in one's confidence judgments of the other's assertions. These results also imply that our version of the representational richness hypothesis (cf., Gill et al., 1998) did not receive full support in that heeding an Actor's arguments *or* confidence judgments did not influence the Observer in a different way than heeding the Actor's argument *and* confidence judgment. However, the representational richness hypothesis received partial support in that heeding *any* additional information (whether it was the Actor's argument, confidence, or both of these) influenced the Observers' level of confidence and resolution.

In addition, and in agreement with the Actor–Observer difference in the level of confidence described above, the Observers' higher confidence when additional information was provided, compared with when it was not, is consistent also with Support theory (e.g., Tversky & Koehler, 1994). This is because both types of additional information, either by themselves or together, might have increased the salience of the answer alternative selected by the Actor and, thereby, increased the Observer's confidence.

The respective contribution of these two explanations to the increase in the Observers' confidence should be analyzed in future research. Moreover, in future research it would be of interest to further investigate the effects of different types of additional information about the Actors' choice process and properties on the realism in Observers' metacognitive judgments.

Frequency judgments. With respect to the Actor–Observer difference in frequency judgments, the results of both experiments of the present study showed significantly lower assessments of one's own actual accuracy compared with that of another person's actual accuracy. These results are in line with previous research (Johansson & Allwood, 2004, Exp. 1 and 2), and are consistent with an interpretation in terms of the pessimistic bias (Klar et al., 1996; Kruger, 1999). In addition to the fact that "egocentrism", as proposed by Kruger (1999), might be an important mechanism behind this effect, there may also be other, perhaps interacting, contributing causes. For example, it may be a consequence of a self-handicapping strategy used in order to keep up self-esteem, and from a more social viewpoint it may also be favorable to elevate, so to speak, one's estimate of another's performance. Our result may also be due to Swedish, or Scandinavian, cultural values according to which one should not pretend to be better than others, including not "showing off" with respect to one's own knowledge. The results of our participants having significantly underestimated their own accuracy (Exp. 1 and 2) while their estimation of the other's accuracy did not differ significantly from perfect realism (Exp. 1) or show overestimation (Exp. 2), are also in line with these interpretations.

As predicted in Exp. 1, the results showed no effect of first making a frequency judgment of one's own and then of the other pair member's accuracy, or vice versa. As noted above, Johansson and Allwood (2004, Exp. 2), in which the two types of frequency judgments were separated by a task involving taking a stand with respect to the correctness of the provided answers, found a clear order effect. Taken together, the results of these two studies suggest that the Actor–Observer difference in realism in frequency judgments depends on the judges' conceptions of the target of their judgments as well as previous frequency judgments. Specifically, and consistent with Mussweiler and Neumann's (2000) idea concerning source attribution, the results of Exp. 1 in the present study suggest that the accessibility of the first frequency judgment might have contributed to a contrast effect in the second frequency judgment (in this case toward a pessimistic bias). However, this suggestion concerning assimilation and contrast effects in the context of the Actor–Observer difference in realism in frequency judgments should be further investigated in future research.

Furthermore, with respect to the confidence–frequency effect, it is striking that in Exp. 1 of the present study the Actor–Observer difference in realism in frequency judgments resulted in a pattern that was quite the opposite to that of the Actor–Observer difference in realism in confidence judgments. While the participants showed higher overconfidence in the item-specific confidence judgments when performing Actors than as Observers, the frequency judgments of their own accuracy were lower than those of the other's accuracy. This finding was partly replicated in Exp. 2, in which no Actor–Observer difference in overconfidence resulted, while the Actor condition resulted in underestimation and lower frequency judgments than the Observer condition, which in turn exhibited overestimation. The discrepancy between these two types of metacognitive judgments provides further support for the assumption that they differ in terms of underlying processes (e.g., Allwood & Granhag, 1996a; Sniezek & Buckley, 1991; Treadwell & Nelson, 1996).

Overall, the results of the present study show that social contextual factors can influence metacognitive judgments in that both confidence and frequency judgments were affected by whether the target evaluated was oneself or another person. Taken together, the almost complete lack of an Actor–Observer difference in the *realism* in confidence judgments, and the apparently robust Actor–Observer difference in frequency judgments, indicates that dissimilar, or even conflicting, levels of realism between these metacognitive judgments can coexist, and that being in the role of an Actor or an Observer can influence these dissimilarities.

Acknowledgment

This research was supported by a grant from The Swedish Research Council (formerly The Swedish Council for Research in the Humanities and the Social Sciences).

Appendix A

$$\text{Calibration} = 1/n \sum_{t=1}^{T} n_t (r_{t\text{m}} - c_t)^2 \tag{A.1}$$

In (A.1), n is the total number of questions answered, T is the number of confidence classes used, c_t is the proportion correct for all items in the confidence class r_t, n_t is the number of times the confidence class r_t was used and $r_{t\text{m}}$ is the mean of the confidence ratings in confidence class r_t.

$$\text{Resolution} = 1/n \sum_{t=1}^{T} n_t (c_t - c)^2 \tag{A.2}$$

In (A.2), c is the proportion of all items for which the correct alternative was selected. A higher value reflects better resolution than a lower value.

References

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology, 68*, 804–825.

Allwood, C. M. (1994). Confidence in own and others' knowledge. *Scandinavian Journal of Psychology, 35*, 198–211.

Allwood, C. M., & Granhag, P. A. (1996a). Considering the knowledge you have: Effects on realism in confidence judgements. *The European Journal of Cognitive Psychology, 8*, 235–256.

Allwood, C. M., & Granhag, P. A. (1996b). The effects of arguments on realism in confidence judgements. *Acta Psychologica, 91*, 99–119.

Allwood, C. M., & Granhag, P. A. (1999). Feelings of confidence and the realism of confidence judgements in everyday life. In P. Juslin & H. Montgomery (Eds.), *Judgement and decision making: Lens-modeling and process tracing approaches* (pp. 123–146). Hillsdale, NJ: Lawrence Erlbaum Press.

Brenner, L., Koehler, D. J., & Rottenstreich, Y. (2002). Remarks on support theory: Recent advances and future directions. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 489–509). New York: Cambridge University Press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 75*, 1–3.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*, 519–527.

Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting, 1*, 155–172.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

Gill, M. J., Swann, W. B., & Silvera, D. H. (1998). On the genesis of confidence. *Journal of Personality and Social Psychology, 75*, 1101–1114.

Granhag, P. A., Strömwall, L. A., & Allwood, C. M. (2000). Effects of reiteration, hindsight bias, and memory on realism in eyewitnesses' confidence. *Applied Cognitive Psychology, 14*, 397–420.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*, 411–435.

Harvey, N., Koehler, D., & Ayton, P. (1997). Judgments of decision effectiveness: Actor–Observer differences in overconfidence. *Organizational Behavior and Human Decision Processes, 70*, 267–282.

Johansson, M., & Allwood, C. M. (2004). *Own-other differences in the realism in metacognitive judgments.* Unpublished manuscript, Department of Psychology, Lund University, Sweden.

Jost, J. T., Kruglanski, A. W., & Nelson, T. O. (1998). Social metacognition: An expansionist review. *Personality and Social Psychology Review, 2*, 137–154.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57*, 226–246.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1038–1052.

Klar, Y., Medding, A., & Sarel, D. (1996). Nonunique invulnerability: Singular versus distributional probabilities and unrealistic optimism in comparative risk judgments. *Organizational Behavior and Human Decision Processes, 67*, 229–245.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*, 461–469.

Koehler, D. J., & Harvey, N. (1997). Confidence judgments by actors and observers. *Journal of Behavioral Decision Making, 10*, 221–242.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 107–118.

Kruger, J. (1999). Lake Wobegon Be Gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77*, 221–232.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121–1134.

Leippe, M. R. (1980). Effects of integrative memorial and cognitive processes on the correspondence of eyewitness accuracy and confidence. *Law and Human Behavior, 4*, 261–274.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art of 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge: Cambridge University Press.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980-1994. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). New York: John Wiley & Sons.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12*, 595–600.

Mussweiler, T., & Neumann, R. (2000). Source of mental contamination: Comparing the effects of self-generated versus externally provided primes. *Journal of Experimental Social Psychology, 36*, 194–206.

Nickerson, R. S. (1999). How we know–and sometimes misjudge–what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125*, 737–759.

Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory, and the eyewitness confidence–accuracy correlation. *Journal of Applied psychology, 81*, 587–594.

Robinson, M. D., Johnson, J. T., & Robertson, D. A. (2000). Process versus content in eyewitness metamemory monitoring. *Journal of Experimental Psychology: Applied, 6*, 207–221.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes, 40*, 193–218.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104*, 406–415.

Sniezek, J. A., & Buckley, T. (1991). Confidence depends on level of aggregation. *Journal of Behavioral Decision Making, 4*, 263–272.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers?. *Acta Psychologica, 47*, 143–148.

Tabachnick, B. G., & Fidell, L. S. (2000). *Using multivariate statistics*. Boston, MA: Allyn and Bacon.

Treadwell, J. R., & Nelson, T. O. (1996). Availability of information and the aggregation of confidence in prior decisions. *Organizational Behavior and Human Decision Processes, 68*, 13–27.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1130.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*, 547–567.

Van Yperen, N. W. (1992). Self-enhancement among major-league soccer players: The role of importance and ambiguity on social comparison behavior. *Journal of Applied Social Psychology, 22*, 1186–1198.

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence–accuracy calibration in face recognition. *Journal of Applied Psychology, 88*, 490–499.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517.